ABSTRAK

Dokumen cetak masih menjadi pilihan beberapa industri untuk menyimpan data-data perusahaan seperti faktur, struk, dan dokumen cetak lainnya. Hal tersebut menimbulkan masalah ketika diperlukan digitalisasi data dari dokumen cetak tersebut. Oleh karena itu, dibutuhkan suatu sistem yang dapat mengkonversi citra dokumen cetak menjadi string agar data tidak perlu dimaukkkan ke komputer secara manual. Saat ini, teknologi yang mampu untuk mengidentifikasi huruf pada citra adalah OCR engine yang didalamnya sudah diprogram untuk melakukan segmentasi, ekstraksi ciri, klasifikasi, training, dan recognition. Salah satu OCR engine yang memiliki akurasi yang paling tinggi (96,38 %) dengan lama pemrosesan paling cepat (4,60 detik) adalah Tesseract. Namun, keakurasian Tesseract bergantung kepada kualitas citra dan noise sehingga diperlukan pengolahan citra tambahan. Oleh kerana itu penelitian ini, dirancang alat pemindai dokumen cetak menggunakan OCR Tesseract dengan tahapan pengolahan citra: grayscaling, unsharp masking, Otsu thresholding, dan dilation dengan library OpenCV. Setelah dilakukan pengujian, Alat mampu mengenali tulisan pada dokumen dengan error 2,58% untuk mengenali kata, error 3,5% untuk mengenali kata dlam suatu kalimat, error 10,5% untuk mengenali kata dalam paragraf, dan error 9,5% untuk mengenali kata dalam dokumen struk. Hasil tersebut untuk ukuran font 16 dengan font Arial, Calibri, Times New Roman, Dot Matrix, dan Fake Receipt.

Kata kunci: *Optical character recognition,* Tesseract, pengolahan citra, OpenCV, pemindai dokumen